A heuristic approach to generate good-quality linked data about hydrography

Luis M. Vilches-Blázquez
Ontology Engineering Group - Departamento de Inteligencia Artificial. Facultad de Informática. UPM
Boadilla del Monte, Madrid (Spain)
lmvilches@fi.upm.es

Oscar Corcho
Ontology Engineering Group - Departamento de Inteligencia Artificial. Facultad de Informática. UPM
Boadilla del Monte, Madrid (Spain)
ocorcho@fi.upm.es

***Abstract* – Current Geographic Information is highly heterogeneous due to the diversity of producers and to their different needs. Geographical databases have different information structures, different levels of abstraction and scale, and are available in different natural languages. This is a major obstacle to overcome when generating good quality Linked Data from these databases. In this paper we describe how we generate Linked Data from heterogeneous hydrographical databases from various Spanish institutions. We provide a characterization of the types of heterogeneity found, based on existing semantic heterogeneity classifications, and describe a heuristic approach to deal with duplicity or co-reference problems.**

Keywords: Geographical databases, heterogeneity, instance duplication, linked data, co-reference, heuristic approach.

I. INTRODUCTION

Linked Data is about employing the RDF language and the HTTP protocol to publish structured data on the Web and to connect data between different data sources, effectively allowing data in one data source to be linked to data in another data source [11]. The focus of this work is to provide Linked Data in the geospatial information context.

Geographical Information (GI) is inherently heterogeneous. This is mainly due to the diversity of information producers, the different scales used to compile this information, and the different treatment and storage of GI as a direct consequence of the different needs and interests of producers.

This heterogeneity is a major problem to overcome when generating linked data for this domain, especially if we are concerned about its quality. One possibility would be to expose the different databases to the linked data world as unrelated entities, leaving the task of linking them to users or other systems. Another possibility would be to carry out information integration processes before the actual generation of the linked data is performed, ensuring a higher quality of the linked data produced.

We will follow the second approach. Particularly, we will be focusing on the detection of instance duplicates in heterogeneous databases, a problem that has been traditionally addressed in the database world. Our aim is to generate higher quality linked data, fixing errors and containing appropriate *owl:sameAs* statements that allow linking the data produced from one database with the data produced from another.

There have been efforts in this direction in the geographical domain. For instance, the concept of geolinked data [12] refers to geographically-related data where the geometry is not directly stored with the attribute data. Instead a geographic identifier is used, which refers to a geometric feature in a separate geospatial data set.

In our work, we provide a characterization of the types of heterogeneity found, based on existing semantic heterogeneity classifications [4, 5], and describe a heuristic approach to deal with duplicity or co-reference problems detected, focusing especially on how to discover (and to ultimately solve) instance duplicates in the information integration process. We illustrate this in the context of several hydrographical databases from various Spanish cartographic institutions.

This paper is organized as follows. Section 2 describes briefly the geographical databases that we will be using to generate our linked data. In section 3, we deal with instance duplication problems in heterogeneous data sources, and one heuristic approach for solving one of these problems is discussed in section 4. In section 5, we provide some results of experimentation. Finally, in section 6, we present some brief conclusions and future work.

II. GEOGRAPHICAL DATABASES' ENVIRONMENT

Traditionally, the process that cartography producers go through to create cartography databases is as follows: first, they identify real world features and give them names; then, they categorize the features and create models, i.e., schemata; finally, they introduce these feature types and their related instances in a database using its underlying syntax [2]. Furthermore, as a consequence of the existence of multiple geospatial producers, it is quite common to find several databases describing, at least partly, the same geographical space. Usually data are collected for specific purposes, and are very different from one source to another [1].

According to [7] GI is increasingly captured, managed and updated with variable levels of granularity, quality and structure by different cartographic agencies. In practice, this approach causes the building up of multiple sets of spatial databases with a great heterogeneity of feature

catalogues and data models. This diversity implies the coexistence of a great variety of sources with different information, structure and semantics and without a general harmonization framework. Besides, this heterogeneity combined with the sharing needs of miscellaneous users and with overlapping information across different sources causes several important problems when we link similar instances to search, retrieve and exploit GI data.

In this work we took into account various Spanish and European geographical information databases. These databases are at different scales (from 1:10 million to 1:200,000) and come from diverse institutions or producers. A common component of these databases is that all sources have hydrographical information related to Spanish geographical feature instances.

With respect to the European databases, we focused on Waterbase, which is used in Water Information System for Europe (WISE). Waterbase is the generic name given to the European Environment Agency's databases on the status and quality of Europe's rivers, lakes, groundwater bodies and transitional, coastal and marine waters, and on the quantity of Europe's water resources. This database has more than 44,000 Spanish hydrographical instances (toponyms) and only contains monolingual information.

On the national database side, we worked with two, which belong to the National Geographic Institute of Spain. On the one hand, The Numerical Cartographic Database, called BCN200 (scale 1:200,000), is a dataset that complies with the required data specifications exploited inside Geographic Information Systems (GIS) environments. This database includes nearly 50,000 toponyms related to hydrographical instances. On the other hand, the National Atlas database (scale 1: 1:500,000) contains information about a collection of maps. This database has more than 1,100 hydrographical instances. Both databases contain multilingual information in the official languages of Spain (Spanish, Galician, Catalan, Basque, and Aranes).

III. INSTANCE DUPLICATION AND LINKED DATA

Instance integration is one of the most complicated tasks in the process of information integration from heterogeneous databases. This is also true in the context of Linked Data, since the explosion in the number of information sources being exposed as RDF has also started to face problems closely related to those addressed in traditional information integration, mainly due to the use of different URIs to identify the same entities. It is often the case that data in different repositories exposed as linked data hold information regarding identical entities, but with different identifiers, what makes information integration and linkage difficult. For example, DBpedia, Geonames, the CIA Factbook and Eurostat all have different URIs for the same country [9].

According to [9, 10] the multiplicity of URIs leads to the problem of co-reference. The co-reference is the problem of ensuring that two different entities do not share the same name or identifier and conversely identifying when two identifiers refer to the same entity. This problem on the Semantic Web can occur in two ways: Firstly, when a single URI identifies more than one resource (e.g.; there are diverse place names (toponyms) with the same name but these toponyms are located at different places) and secondly when multiple URIs identify the same resource (e.g.; Spain has different URIs which depend on source. Thus, in DBpedia appears as 'http://dbpedia.org/resource/Spain', while in Geonames has the URI 'http://sws.geonames.org/2510769').

The standard way of dealing with a set of URIs which refer to the same resource is to use *owl:sameAs* property to link between them. The semantics of this property means that all the URIs linked (resources) with this predicate have the same identity [13], that is, the subject and object must be the same resource. The major disadvantage with this approach is that two URIs become indistinguishable even though they may refer to different entities according to the context in which they are used [9].

On the Semantic Geospatial Web this presents a problem when there is a need to link together knowledge compiled within diverse databases from disparate information providers.

Within the database community the problem of co-reference is referred to as record linkage. The need for record linkage arises when records or files from different databases need to be joined or merged. Each database could have duplicate records of the same person or thing which, when amalgamated, would make the data inconsistent or "dirty". Next, we describe different problems related to the duplication of instances found in diverse comparison processes in heterogeneous hydrographical data sources.

A. Instance duplication problems

With the Linked Data initiative in its early stages, it seems difficult to think about problems that appear when we attempt to integrate different and heterogeneous data sources. As an initial step we have analyzed and found out different instance duplicate problems in several available sources. These problems are classified following the layered approach (lexical, syntactic, and pragmatic) that commonly appears in works that deal with semantic interoperability [4, 5]. Next, we provide a non-exhaustive list of instance duplicate problems.

- The lexical layer. This layer deals with the ability to segment the representation into characters and words (or symbols) [4]. In this layer, we have observed diverse types of problems at different analysis levels. With regard to problems in a single database, we have found a problem type related to Spanish signs, that is, the differences between instances due to the presence or absence of accent and special letters, such as, ñ, ç, ª, º, etc. This problem appears when we check one or various databases at the same or different scale.

- The syntactic layer. This layer deals with the ability to structure the representation in structured sentences, formulas or assertions [4]. In this layer we checked some important errors related to typographical mistakes in BCN200, which are related to recording errors in [6]. For instance, we found errors such as "S?quia" ("*Acequia*"), "Aigua" ("*Agua*") or "Braz" ("*Brazo*") and others. Finally, another problem found quite frequently is that of similarity of instance names, i.e., similarity between different strings (e.g. "*Perrera, Arroyo de la*" (Perrera, Stream) and "*Perreras, Arroyo de las*" (Perreras, Stream)).
- The semantic layer. This layer deals with the ability to construct the propositional meaning of the representation [4]. In this layer, in a single database context, we highlight as an important problem that of different names of feature types that is, an instance name which is associated with different geographical feature types, regardless of its spatial location and of the presence or absence of a relationship between these geographical features, for instance, "*Arroyo de Periquito*" (Periquito Stream) and "Rambla Periquito" (Periquito watercourse). On the other hand, in the different databases integration context at the same scale we find problems related to different classification or viewpoint (e.g. "*Arroyo de las Rozas*, *Dehesa*" (Rozas Stream, Meadow) and "*Arroyo de las Rozas*" (Rozas Stream)). In the first example, the National Atlas database classifies this instance as a place, whereas the second instance is classified as a stream in BCN200. This problem is related to cognitive heterogeneity [2].
- The pragmatic layer. This layer deals with the ability to construct the pragmatic meaning of the representation (or its meaning in context) [5]. In a single database we found problems related to official or alternative names, which can be subdivided into two types: (1) Official or alternative names that can be found at different fields of a database table (for instance, "*RÍA DE ORTIGUEIRA*" and "*Ría de Ortigueira e Ladrido*" in BCN200). (2) Official or alternative names that can be found at different tables of a database (e.g. "ARETA" and "URRAUL"). With respect to the problems related to different names of the feature types which were treated above, we have subdivided them into two subtypes and these are related to cartographic representation, scale factor, and conceptual abstraction; such problems are related to generalization and aggregation, as was exposed in [3]. There are database instances that share the same instance names (for instance, "*Acequia de la Fábrica de Luz"* and "*Fábrica de Luz, Caz de la*"), but their related feature type is a superclass of another feature type; for instance, channel is a superclass of irrigation channel. The previous instances appear in BCN200 and in the National Atlas database, respectively. We must add that there are some instances which have the same instance name, but their related feature type is a subclassOf another feature type, e.g.; "*Riachuelo de la Cañada*" (Cañada Creek) is subclass of "*Arroyo de la Cañada*" (Cañada Stream). Finally, we also found a problem related to duplicity or co-reference (the same instance names appear duplicated within different data sources). This information duplicity is caused by the diversity of producers and datasets.

IV. HEURISTIC APPROACH FOR INSTANCE DISAMBIGUATION

Considering the aforementioned problems and the difficulties to disambiguate instances in the hydrographical domain, a combination of techniques should be applied.

In our approach we consider that not only instance labels are relevant for instance disambiguation but also context has a key role in this process. That is, algorithms for instance disambiguation must take domain knowledge into account. For that reason, we follow a heuristic approach that combines different levels of domain (in)dependency: hydrographical (e.g., river is similar to a water stream, but quite different from a lake) geographical (a line and polyline can refer to the same entity when considered at different scales, whereas a line and a point should be considered different), and domain independent aspects (entities with the same name could be considered initially similar to each other).

In our case, we use an OWL hydrographic domain ontology, called *hydrOntology* [8] as one of the main knowledge resources used in our heuristics. We also take into account some general characteristics associated with geographical resources. As an example, we describe one of this heuristics in section 4.1.

Once instance duplication problems are solved we have to proceed to the final generation of the linked data to be published. We claim that not every piece of data that has been used in the instance disambiguation process has to be included in this final code, because it may not be relevant for external use. Domain independent data like name, length, surface, coordinates and so on are included, whereas the scale factor and geometrical representation should probably not be included. Nevertheless, this is out of the scope of this paper.

A. *A sample heuristic for detecting river duplicates*

This section describes one of the heuristics that we have implemented in our linked data generation system. This heuristic solves duplicity or co-reference problems that belong to the pragmatic layer described in section 3, taking into account the aforementioned domain dependent and independent considerations.

Let us consider the set of databases described in section 2. We have different URIs for the same resource. For instance, next we show six different ones that belong to the Ebro River.

http://ign.fomento.es/hidrografia/BTN25/Río_Ebro
http://ign.fomento.es/hidrografia/BCN200/Río_Ebro

http://ign.fomento.es/hidrografia/NOMGEO/H25/EbroRio
http://ign.fomento.es/hidrografia/ANE/Río_Ebro
http://EuroGlobalMap/WatrcrsA/BH502/RioEbro
http://chebro.es/RIOS/1491

In order to disambiguate these different URIs (their associated instance names) and generate a harmonized view of hydrographical information we take into account the following independent and dependent characteristics related to our knowledge domain. With respect to domain independent characteristics we compare different instance names (URI) within each database (*Db*). The implemented algorithm detects and counts identical and very close strings (labels and related feature types, e.g.; "Ebro"+ "River").

Given the set £ of all labels in Db1 ∪ Db2 … ∪ Dbn
And the set F of all features in Db1 ∪ Db2 … ∪ Dbn
ForAll ℓ ϵ £ do:
labelQuantity[ℓ] = |{x ϵ Db1 ∪ Db2 … ∪ Dbn / x.label = ℓ}|
ForAll ℓ ϵ £ and f ϵ F:
labelQuantity[ℓ, f] = |{x ϵ Db1 ∪ Db2…∪ Dbn / x.label = ℓ ∧ x ϵ f}|

Moreover, we also compare semi-automatically different values of length attribute assigned to each instance within every database. In this comparison process a domain expert has to select length attribute from different databases (*e.g.; db1.length, db2.riverLength, db2.length, db3.totalLength*). Here, we specify a threshold with a similarity score of 10 % on length values of sample. Non-compliant values are considered as different information, which should be reviewed by an expert.

$$Compare\ \forall instance.label(length) \in Db_1\ with\ \forall instance.label(length) \in Db_{2,3,n}$$

Another attribute used in the disambiguation process are coordinates, that is, we compare initial (X) and final (Y) points of each instance in the real world. Previously, we carry out an on-the-fly coordinate transformation process (UTM/Geographical) to harmonize this information, if it is necessary. In this comparison process we establish a similarity threshold around 25% of geographic distance with instances' coordinates belonging to the sample. Non-compliant values are considered as different information.

$$Compare\ \forall instance.label(coordinates.(x),(y)) \in Db_1\ with\ \forall instance.label(coordinates.(x),(y)) \in Db_{2,3,n}$$

Finally, with respect to independent characteristics, we take into account other resources to enrich the disambiguation process. Thus, we use *hydrOntology* [8] to set relationships with other associated hydrographical features (e.g.*IsMadeUpOfReservoir*, *hasTributary*, or *belongToBasin*) and also use specified knowledge about this domain (cardinality, disjoints, axioms, etc.) in the ontology. The connection between different instances and diverse relationship types of the ontology are not decisive, although it can be binding in the disambiguation process.

With respect to domain dependent characteristics (geographical issues) we consider information related to scale and geometrical characteristics to each instance. The goal of this information is to add a consistency checking process for the disambiguated instance.

With respect to geometry, we collect automatically the representation type (point, line, and surface) in each spatial database from coordinates set related to each instance. This geometry comparison is not decisive, although it can be binding in the disambiguation process.

$$Compare\ \forall instance.label(geometryType) \in Db_1\ with\ \forall instance.label(geometryType) \in Db_{2,3,n}$$

Besides, we consider scale information, which is associated manually to each database and, therefore, each hydrographical instance.

Finally, a domain expert checks results of comparison process and appropriate to them we manually establish relationships between different identical instances through *owl:sameAs* property.

V. EXPERIMENTATION AND DISCUSSION

In order to check our heuristic approach and databases' quality, firstly, a domain expert selects a sample of main Spanish rivers (100 instances) from domain references. This sample is used to establish the gold standard, that is, a pre-defined dataset by this expert as correct information. Later we use databases aforementioned to compare, check similarity, and disambiguate information through our heuristic approach. In Table 1 we show some statistical details related to percentages of identical information between controlled databases and the sample.

Table 1. The evaluation of our heuristic approach

	Waterbase	**National Atlas**	**BCN200**
Name	60%	100%	83%
Length	13%	85%	34%
Coordinates	70%	-	75%
hydrOntology			
IsMadeUpOf Reservoir	-	95%	75%
hasTributary	47%	100%	-
belongToBasin	60%	100%	-
Geometry	line	line	line
Scale	1:250,000 to 1:10 million	1: 500,000	1: 200,000

With respect to domain independent characteristics, in the comparison process of strings, we highlight the amount

of identical instance names detected, especially in National Atlas database (100%). Nevertheless, length attribute similarity between previous instances is not large in Waterbase and BCN200, though in National Atlas is 85%. As a consequence of these results, we assume that this attribute does not represent the right length of rivers in the real world into Waterbase and BCN200, but this attribute has information on geometrical elements. Regarding coordinates, the results show around 70% identical (strings) instances are contained within the established similarity threshold.

With respect to relationships with other associated hydrographical features, we can see that National Atlas database shows high percentages of similarity, while BCN200 only has information about associated reservoirs. With respect to domain dependent characteristics, all databases represent rivers by means of lines. Finally, these databases have information at European, National (1:500,000) and province (1:200,000) scale, respectively. Therefore, these sources have different granularity.

These results illustrate that our heuristic approach takes into account key information in the detection and disambiguation process of duplicated instances related to rivers. Moreover, these results highlight quality and reliability of National Atlas database.

VI. CONCLUSIONS AND FUTURE WORK

We have provided a brief description of geographical databases' environment. Furthermore, in the context of linked data and information integration of heterogeneous hydrographical databases, we have detected a large number of instance duplicate problems at the semantic heterogeneity level. In order to contextualize these problems we have employed different heterogeneity types. Finally, we raise one heuristic approach to solve duplicity or co-reference problems.

Regarding future work, we aim to create an exhaustive list of instance duplicate problems in geographical databases, hence extending our work on hydrographical databases. For that reason, we will continue working with the problems detected previously, through the formalization and implementation of more heuristic approaches.

ACKNOWLEDGMENTS

This work has been partially supported by the R&D project España Virtual, funded by Centro Nacional de Información Geográfica and CDTI under the R&D programme Ingenio 2010, and the Spanish R&D project "GeoBuddies" (TSI2007-65677C02).

REFERENCES

[1] Devogele, T., Parent, C., Spaccapietra, S. (1998) On spatial database integration. International Journal of Geographical Information Science, vol 12 - Issue 4, pp. 335-352.

[2] Bishr Y (1998) "Overcoming the semantic and other barriers to GIS interoperability". International Journal of Geographical Information Science, 12(4): 299–314.

[3] Naiman, C. F., Ouksel, A. M. (1995) A classification of semantic conflicts in heterogeneous database systems. Journal of Organizational Computing, Volume 5, Issue 2, pp. 167 - 193.

[4] Euzenat, J., 2001. Towards a principled approach to semantic interoperability. In: Gómez-Pérez A., Grüninger M, Stuckenschmidt H, Uschold M. (eds.) IJCAI 2001 Workshop on Ontology and Information Sharing, Seattle, Washington.

[5] Corcho O. (2005) A layered declarative approach to ontology translation with knowledge preservation. Frontiers in AI and its Applications. Dissertations in AI. IOS Press.

[6] Chatterjee A., Segev A., (1995) "Rule Based Joins in Heterogeneous Databases," Decision Support Systems, vol. 13, no. 1, pp. 313-333.

[7] Gómez-Pérez A, Ramos JA, Rodríguez-Pascual A, Vilches-Blázquez LM (2008) The IGN-E Case: Integrating through a hidden ontology. In Headway in Spatial Data Handling. 13th SDH'08. Lecture Notes in Geoinformation and Cartography, Ruas, A., Gold, C. (Eds.) pp. 417-435. Montpellier, France.

[8] Vilches-Blázquez LM, Bernabé-Poveda MA, Suárez-Figueroa MC, Gómez-Pérez A, Rodríguez-Pascual AF (2007) "Towntology & hydrOntology: Relationship between Urban and Hydrographic Features in the Geographic Information Domain". In Ontologies for Urban Development. Studies in Computational Intelligence, vol. 61, pp. 73–84. Springer.

[9] Jaffri, A., Glaser, H., and Millard, I. (2007) URI Identity Management for Semantic Web Data Integration and Linkage. In Proceedings of the Workshop on Scalable Semantic Web Systems (Vilamoura, Portugal) Springer.

[10] Jaffri, A., Glaser, H. and Millard, I. (2008) URI Disambiguation in the Context of Linked Data. In : Linked Data on the Web (LDOW2008), Beijing, China.

[11] Bizer, C., Heath, T., Idehen, K., Berners-Lee, T. (2008) - LDOW2008. In WWW'08 pp. 1265-1266, Beijing, China.

[12] OGC (2004) Geolinked Data Access Service (GDAS), Version: 0.9.1, OGC Inc. Wayland, MA, USA.

[13] Bechofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., Mcguiness, D.L., Schneider, P.F. and Stein, L.A.OWL Web Ontology Language Reference, Technical Report, W3C, http://www.w3.org/TR/owl-ref/